FEATUR.UX: An Approach to Leveraging Multitrack Information for Artistic Music Visualization

Ireti Olowe
Center for Digital Music[†]
Queen Mary
University of London
i.o.olowe@qmul.ac.uk

Mathieu Barthet[†]
m.barthet

Mick Grierson
EAVI Group, Depart. of Computing
Goldsmiths College
University of London
m.grierson@gold.ac.uk

Nick Bryan-Kinns[†]
n.bryan-kinns

ABSTRACT

FEATUR.UX (Feature - ous) is an audio visualisation tool, currently in the process of development, which proposes to introduce a new approach to sound visualisation using pre-mixed, independent multitracks and audio feature extraction. Sound visualisation is usually performed using a mixed mono or stereo track of audio. Audio feature extraction is commonly used in the field of music information retrieval to create search and recommendation systems for large music databases rather than generating live visualisations. Visualizing multitrack audio circumvents problems related to the source separation of mixed audio signals and presents an opportunity to examine interdependent relationships within and between separate streams of music. This novel approach to sound visualisation aims to provide an enhanced listening experience in a use case that employs non-tonal, non-notated forms of electronic music. Findings from prior research studies focused on live performance and preliminary quantitative results from a user survey have provided the basis from which to develop a prototype for an iterative design study that examines the impact of using multitrack audio and audio feature extraction within sound visualisation practice.

1. INTRODUCTION

Sound visualisation is primarily performed using mixed tracks of mono or stereo audio. One type of sound visualisation created for live performance features representations of the score, which exhibit temporal and tonal structures of music that complement the listening experience using an archetype of representation that elucidates the relationship between music perception and the musical staff. These visualisations commemorate the traditional notated form of written music and are not derived from characteristics of inherent signal properties, capturing performers' expression. Other examples of sound visualisation, which utilise the signal properties of audio signals to explore synchronization between sound and image, rely on databases of pre- and post- processed video clips and loops, and use a limited selection of common high-level audio feature extractors: e.g. tempo expressed in beats per minute (bpm), sound intensity, pitch, and timbre.

FEATURE.UX is a visualisation tool that allows users to create generative visualisations by connecting directed paths between nodes signifiers that represent an audio signal or audio feature, graphic methods of drawing and, color and threshold parameters within the primary workspace of a graphic user interface (GUI). Each completed path creates a separate, layered visual composition, which is exported to its own screen buffer to be displayed on a monitor or projected. Modular panels afford the user the flexibility to organise the interface within the workspace. Data visualisations monitor the behaviour of audio features in real time to indicate how mapped visuals respond to an audio signal. The ability to work with many instances of an audio signal or feature object provides users the opportunity to have greater and finer access to the sonic material, determine the complexity of layered visual compositions and to experiment with mapping strategies between audio features and visual properties.

2. BACKGROUND

2.1 Related Work

Applications of multitrack sound visualisation in literature are documented within the fields of data sonification, audiovisuals and sound visualisation.

Diniz, Demey and Leman modeled complex event structures in electronic music to develop a system for multilevel sonification of data [1]. Song and Beilharz explored perceptual musical characteristics such as timbre to identify aesthetic considerations during the sonification of multiple data streams through visual spatialisation [2].

Audiovisual systems such as the Reactable [3] and AV Clash [4] can be considered multitrack visualisation tools. Both visualize interdependent relationships between multiple, simultaneous streams of sound within a performance context. In contrast to visualisations that produce a visual representation of sonic information from an audio signal, these projects feature individual sound objects, which generate visuals that also synthesize the accompanying audio component of the performance. Relevant comparative models to our proposed FEATURE.UX system are those whose graphics and aesthetics are generated from the signal analysis of audio but whose visual components are not also tasked with providing a simultaneous audio accompaniment.

Throughout multitrack audio research, audio streams have been visualized to provide analysis and control. Hiraga, Mizaki and Fujishiro developed a system to review live performance as a method for performers to share an ensemble experience between co-players, compare different degrees of expression between performances and to understand the intent and mood of each performer [5].

Dahyot, Kelly and Kearney also explored the use of multitrack stream visualisation for live performance. Their objective was to utilise separate streams of data to control the environment in which artists performed. An individual stream of audio output from each band member is used to trigger lighting events and enable animations [6].

Soma is an applicable visualisation tool designed to offer real-time multitrack visualisations. The system renders visuals from separate streams of MIDI data for live performance. Bergstrom developed Soma to challenge typical mapping conventions of limited high-level features used to define the sound-image relationship. Bergstrom wanted to exceed the limited conventions of visualisation practice that mapped visuals to the beat or amplitude of music. He proposed to gain deeper access to lower level audio features to explore the elements of expectation: tension and surprise [7]. His system enables a group or an individual performer to display visual music generated from the output of performed instruments [8]. The visuals in Soma map to MIDI parameters: i.e. scales, notes, chords, tempo, volume and force.

Another aim of Soma was to create intuitive control interfaces to replace the ubiquitous use of knobs, buttons and sliders. The system is composed of a graphics rendering module, a module to monitor gestures and control input and a module to manage mapping between the renderer and control information. Real-time graphics are produced through the interpretation of MIDI sent via Open Sound Control (OSC) or multi-channel musical data processed using visual synths that produced unique effects [9].

The decision to forgo hardwired mapping between audition and visual domains allows a performer to generate dynamic graphics throughout the visual performance. Soma separates the role of the musician from the role of the visual mix engineer. While information is generated with played instruments, the visual engineer performs by deciding how to create links between the data and the visual synths [9].

2.2 Multitrack Audio

It is our hypothesis that a multitrack approach allows the generated visualisations to feed upon a richer, more abundant source of data to produce a more complete representation of its characteristics and expression.

A major difference between using a single track of mixed or multiple, pre-mixed tracks of audio as the data source from which to generate visualisations lies in the amount of available, accessible and employable information. Mixing and mastering individual stems to produce a final mix may significantly alter the fidelity of audio features, depending on the feature extracted from the pre-mixed tracks. In cases where the processes fail to preserve the sonic distinction between audio tracks, particular traits of the sonic information within individual stems may be lost. The benefit of multitrack visualisation offers a richer pool of data from which to extract musical features, map parameters and exhibit their behaviours and relationships in the visual domain. Fazekas demonstrates that tracks analyzed independently impart information that would be nullified by the mixing process [10]. As stated by Hargreaves, independently analyzed tracks avoid occlusion within salient portions of audio in individual tracks that become difficult to isolate and analyze in mixed signals [11].

Software tools for audio and composition analysis have employed multitrack audio. TaCEM was developed to study the influences of technology on electroacoustic music composition. [1] Coupries EAnalysis framework sought to introduce new composition tools through the exploration of graphic representation. [2] Providing support for multitrack audio in creative software is gaining popularity. VDMX [3] routes audio signals over OSC from Ableton Live using Soundflower. Magic Music Visuals [4] also supports multitrack audio. Other popular software packages such as Quartz Composer, [5] Pure Datas Gem, [6] Max/MSPs Jitter, [7] Touch Designer, [8] and VVVV, [9] which allow users to fabricate their own tools by patching modular objects together can also support multitrack audio if assembled by their user.

Native Instruments, a manufacturer of hardware and software audio production and performance instruments, has developed and introduced the Stem, [10] an open sourced, audio file format built upon the MP4 framework. The Stem file format incorporates five separate audio stems. A stem is an independent track of audio that may be mixed with additional stems during the production of music to compose a mixed and mastered mono, stereo or multichannel audio file. Each of four stems within the Stem format file holds one dedicated stream of audio, e.g. drums, vocals, bass, harmony. The fifth stem holds the original stereo master of the mixed composition. This file format lets one independently interact, modify, and, isolate or combine playback of any one or more streams of audio in real time. We plan to support the Stem format in future iterations of FEATURE.UX.

2.3 Feature Extraction

Feature extraction can operate on a time-varying or steady audio signal. A signal is partitioned into shorter segments during which a signal can be considered to be locally invariant. The representation of an audio signal can be extracted from within the time domain directly from the wave-

[1] http://www.hud.ac.uk/research/researchcentres/tacem

[2] http://logiciels.pierrecouprie.fr

[3] https://vidvox.net/

[4] https://magicmusicvisuals.com/

[5] http://quartzcomposer.com/

[6] https://puredata.info/

[7] https://cycling74.com/

[8] http://www.derivative.ca/

[9] https://vvvv.org/

[10] http://www.stems-music.com/

form or after the signal has been transformed into the frequency domain to disclose its spectral characteristics [12]. An audio signal, analyzed either locally by frame or globally over longer durations of time, reveals structural or semantic properties descriptive keywords defined from familiar language used to describe their sonic characteristics [13] by which they can be classified and understood [14].

Common audio features, derived from music composed of pitched sound objects of short duration and fixed timbre organised into larger structures are identifiable and quantifiable [15]. Wisharts quote categorises properties of classical, contemporary and popular music (the current emergence of electronically-tinged popular music, notwithstanding) whose formulaic, melodic and harmonious arrangements are constructed from phrase structures that constitute sonic events [16] [17]. The task of creating a representational visualisation of audio from music of this specification despite its complexity can be accomplished directly from the data, where known values can be extracted from the notation.

Music whose characteristics are exhibited through complex textures and shifting, evolving transformations rather than notes and chords are encompassed in what Edgard Varse coined organized sound in the 1920s [18]. This description has since evolved into a class of music that consists of many forms, structures and styles. [11] Electronic music is a variegated signifier that endeavors to describe the diverse methods of composition and aesthetics of all encompassed sub-genres it aims to define [17] [19]. The type of music with which this research is concerned is non-notated spectral music, which fails to provide neat numerical data by which to showcase its attributes no note or MIDI information may exist to appraise the contour of an envelope or detect the discrete distinct grains within layered scales of sonic or temporal structures [20]. Music Information Retrieval and audio content-based processing can help close the gap towards extracting musical and perceptually-relevant information from a non-notated, electronic music audio signal [21].

Dahan outlines problems with using computer analysis on electroacoustic music signals; tools are primarily developed to evaluate Western, tonal and pitched music. He suggests three MIR techniques that can be used to analyze electronic music signals. Firstly, using Mel-Frequency Cepstral Coefficients (MFCCs) produces perceptually-relevant impressions of timbre. Secondly, signal analysis of electronic music can take advantage of MIR pattern recognition techniques used to access audio repetition, which can be viewed as a trademark of certain electronic music genres, e.g. chiptune, dubstep, breakbeat and glitch. Also present in traditional music, repetition, presents itself in irregular patterns that can be appear within various time scales [20]. Finally, the segregation of sound may not be an obstacle as multitrack audio sources remain distinct [22].

2.4 Audio Visual Practice, Performance and Tools

Baker et al. collated writings from blog posts that contributed to an online community of 13 writers and 19 commenters [23]. Built around real-time live performance and media, the community discussed topics associated with real-time media from the perspective of the performer, performance and audience over a period of three months. Posts from the project expressed that its writers want to be introduced to new methods of performance, i.e. shared performances that break the limitations of a solo VJ presenter [23]. Comments suggested that VJs should have increased prominence when working with DJs who benefit from more notoriety and visibility [24]. Views expressed that the experience of the viewer should be less dependent upon their interpretation of the VJs artistic intent and that the causal relationship between sound and image should be easily discernible. The experience of live performance should be presented with new representations of time and innovative aural, visual and spatial aesthetic experiences [24].

Hook et al. analyzed the expressive interactions of VJs. The research team filmed a documentary with four VJ collectives as they transitioned between practice, preparation and performance. They hosted focus groups with the VJs and asked them to re-edit the documentary according to specific topics. The findings of the experiment were categorised by themes: the aspirational category focused on artistic intent, goals and desired outcomes for their performances; the interaction theme concentrated on the impact of interaction upon the VJs practice; and the live category addressed the significance of liveness that the VJs placed on their practice. Results indicated that the VJs want live visual performance to evolve and become an integral part of musical performance. They want software that facilitates visual improvisation, mutability and that is less dependent on rendering assets. The artists sought to obtain finer control of the audio, interact with the data and receive immediate feedback from their actions. The VJs expressed the need for flexible, reconfigurable GUIs and tools that influence creativity by affording fewer options. While the VJs would like to engage the audience by revealing the causal link between sound and image, they also expressed that they would like to constrain audience interaction [25].

Correia and Tanaka surveyed the landscape of existing computer-generative tools for audiovisual performance and composition. Taking a user-centered approach, they conducted interviews and hosted workshops focused on the expression and usability of tools, and audience engagement. Participants called for modular GUIs; integration of tools across a variety of software; and tools that afford greater expressibility, generative capabilities, flexible timelines, and the ability to expose the performers effort to the audience [26].

In VJ: Audio Visual Art + VJ Culture, author and artist, Faulkner, also known as D-Fuse, provided a passport to VJing by providing a thorough survey of audiovisual cultures, artists and resources from around the world. In his remarks, the artist calls for new methods of visualisation, perceptible expressions of structural relationships between the aural and visual domains in addition to aesthetics, the ability for the user to influence the audio portion of the VJ performance, and format agnostic software. He proposes a VJ practice that is less technical and reliant on amassed

[11] http://ears.pierrecouprie.fr/spip.php?rubrique3

collections of video assets, decreases the pre-production required to manage the assets, and that breaks away from looped-video presentations [27].

3. METHOD

3.1 User Survey

A combination of iterative design and thematic analysis will be used to evaluate FEATURE.UX as a tool to control parameterized graphics using audio features extracted from multitracks. Other areas for investigation include expressibility of the user, mapping between audio features and visual parameters, and usability.

The attributes of FEATURE.UX were selected after performing a survey of research literature focused around VJ tools, practice and theory and, live performance and interfaces. Also, a comprehensive survey was designed to collect information from live visualists, sound visualisation and audio-visual artists, and VJs. 31 open-ended questions were focused along six categories: experience, preparation, performance, mapping between audio and visual domains, multitrack audio and technological enhancement. Thirty opinion scale/Likert questions focused on multitrack sound visualisation, audio feature extraction and applied mapping strategies used to link sound with image. Demographic questions included queries about the subjects work samples for context. The online survey was hosted on the BOS online survey platform run by the University of Bristol and distributed via email, Twitter, Facebook and throughout several audiovisual communities, digital artist networks and commercial software forums. Quantitative results were evaluated from the responses of 22 participants (three women and 19 men aged from 21 to 57 years submitted surveys. The median age is 36.5 years.) Fifteen (68.2%) participants are professionals, six (27.3%) are amateur performers and one (4.5%) is a hobbyist.

3.1.1 Audio Stem Results

Overall, the findings show that the practice of using multitrack audio to create visuals is not prevalent enough to assess. More artists (27.2%) would use audio stems to create their visuals if given access than the 18.1% who stated they would not. But, 49.5% of the participants are at least satisfied with only having access to the stereo or mono mix to create visuals, 22.7% either felt indifferent or disinterested, and the same percentage, 22.7%, were not satisfied with having access to only a stereo or mono mix.

For 45.5% of the respondents, access to stems would provide greater control of the audio data with which to create visuals as shown in Fig. 1. But, the same 45.5% acknowledged that including stems within their workflow would make the process more complicated, some noting that the added complexity may not be worth the effort.

The availability of stems is not directly linked to the assessed quality of the finished composition. 50.0% felt that working with stems wouldnt necessarily make their visuals better, however, 18.2% agreed and 18.2% disagreed that employing them would make their visuals more meaningful. Also, 31.8% believed that having access to stems would create a stronger link between sound and image.

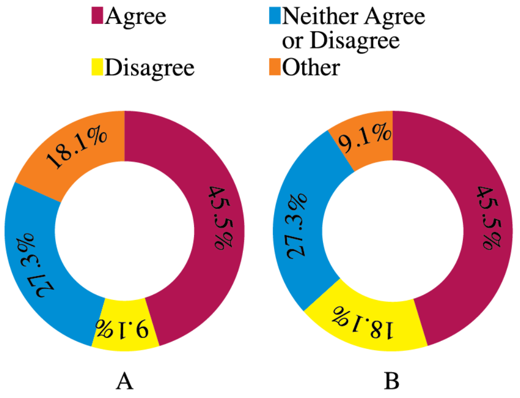

Figure 1. Chart A shows the percentage of participants who felt that having access to audio stems to create visuals would afford more control over the musical data. Chart B shows how they responded to adding complexity to their workflow by employing stems to create visuals.

Statements expressed concern that utilizing stems would affect audience engagement. 36.4% participants agreed and 36.4% disagreed that using stems to create visuals may render them too complicated for an audience to follow. Additionally, one artist (4.5%) offered the opinion that the assessment of complexity for particular visuals depends on the application.

3.1.2 Audio Feature Results

45.4% felt that additional audio features beyond what is already available are not needed to create visuals and 40.9% responded that the beat and the pitch are the most meaningful features to use. However, 59.1% neither agreed or disagreed, which may signify a deficiency of awareness about audio feature extraction as a tool for visualisation. Written responses stated that the lyrics, style of music and the extracted spectrum can be equally important.

Only 9.1% of the participants felt strongly uninterested in using additional features to create visuals. Although 45.4% of participants deemed access to more features unnecessary, 50.0% of the artists were at least interested in having access to more features as shown in Fig. 2.

Even though the results suggest that using multitracks would adversely affect audience engagement, 81.8% of the artists are not concerned that viewers may not understand the visuals created using additional audio features. The artists (40.9%) are more concerned with the synchronization between audio and visuals. 22.7% agreed and 22.7% disagreed that additional features will help create a tighter sync between sound and visuals.

3.1.3 Mapping Results

The results imply that the practice of mapping is more of an intuitive exercise than based on an exact system as shown in Fig. 3. 22.7% of artists felt that creating arbitrary pairings between audio features and visual parameters is an adequate method to create a link between the two domains. 36.4% disagreed. Furthermore, there is no consensus that

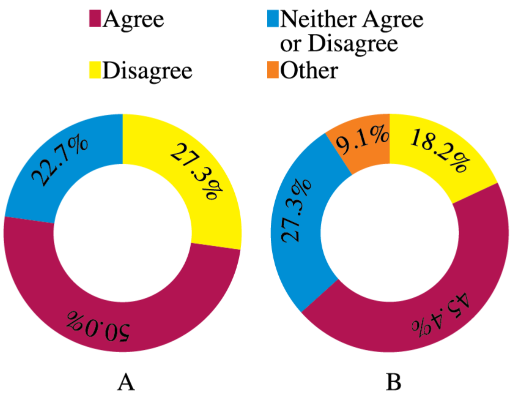

Figure 2. Artists expressed interest in having access to additional audio features with which to create their visuals as shown in Chart A. Although, as shown in Chart B, the participants overwhelmingly agreed that using additional audio features to create visuals is not essential to their practice.

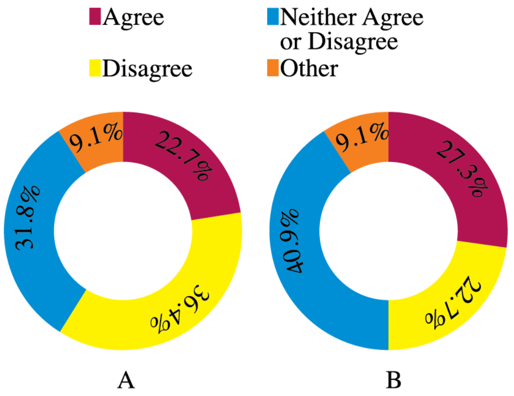

Figure 3. In Chart A, the artists assess mapping between audio parameters and visual attributes as a random exercise within their practice. Whether there exist established conventions between audio parameters and visual attributes is measured in Chart B.

there either are or arent established conventions between pairings, 27.3% agreed, 22.7% disagreed. Most (40.9%) responded neutrally. It follows that a majority of artists (72.7%) create their own rules when mapping between audio features and visual parameters.

Establishing a causal relationship between sound and image is a concern for 45.4% of the artists. 100.0% of the participants disagreed that an audience could only determine links between sound and image if the audio and visuals were mapped using a 1-to-1 mapping strategy, in which parameters in one domain are linked to one and only one feature in the other [28].

The degree of expression or meaning that an artist seeks to achieve within the visuals does not seem to be directly related to the implemented mapping strategies. Either a 1-to-many or many-to-1 relationship between sound and visuals allows for one parameter in one domain to be represented by more than one parameter in the other. Implementing one of these mapping strategies is more likely to increase the complexity of a composition since links between associated parameters are superimposed [28]. However, there is no consensus about the visible influence of executing these strategies. 68.2% neither agreed or disagreed that the link between sound and visuals can be expressed or distorted due to the employed mapping between the two and, 13.6% equally agreed and disagreed.

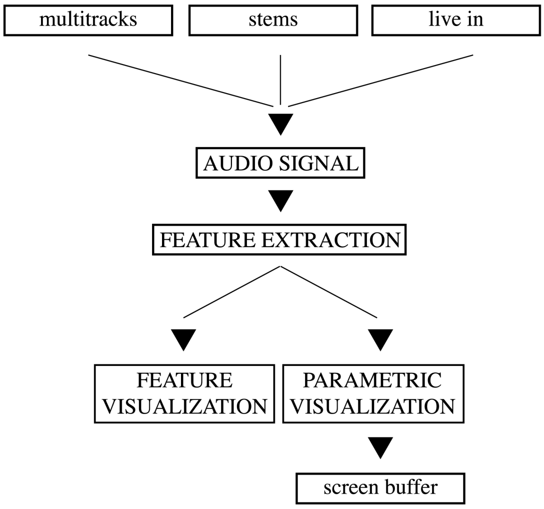

Figure 4. Schematic of the FEATUR.UX system.

Despite the mapping strategy employed, 68.1% of the participants disagreed that audiences should be able to relate every sound event in the music with an accompanying visual. In addition, the viewer's interpretation, which may depend on the type and location of a performance [24], is not expected. 36.6% felt that the audience does not have to understand the visuals, even though 40.9% are interested in conveying meaning through their performances to those who experience them.

3.1.4 Quantitative Results Discussion

Most artists neither agreed or disagreed with 26 of the 30 Likert statements. Utilizing stems from multitrack audio and additional, uncommon audio features to create live visuals is not yet popular enough to build opinions about their impact on practice and performance. In addition, there is no established common language to distinguish between disciplines within live audiovisual practice and performance. Carvalho and Lund sampled the live audiovisual community to learn how practitioners define their own practice. The results of the 2014 international survey found that the boundaries and language used to define practices within visual music, expanded cinema, live cinema, VJing and audiovisual performance are continuously debated, fluid and ambiguous. Finding consensus about the practices involved within live audiovisual performance using terms like visuals and visualisation is difficult when their meanings are malleable and depend on their application, usage and context [24].

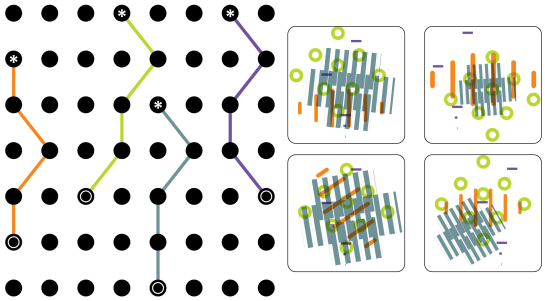

Figure 5. A representation of sound visualization within FEATUR.UX. (A) Directed paths are drawn within the workspace. The paths, which flow from top to bottom, begin with an audio signal pin (noted by an asterisk) and end with an output window pin (noted by a white circle). (B) The illustrations to the right of the directed paths represent consecutive frames of a visualization.

3.2 FEATUR.UX Prototype

The FEATUR.UX prototype is programmed using openFrameworks. [12] The ofxGui addon was used to build the user interface. Maximillian is an audio synthesis and signal processing library written in C++, whose addon, ofxMaxim, was chosen for its native real-time audio analysis and feature extraction capabilities. It is an easy-to-use framework with a collection of fundamental feature extractors commonly used for music information retrieval [29].

A tabulated list of desires and needs articulated by participants to evolve the practice and performance of live visuals, as mentioned in previous research discussed in Section 2.4, was used to choose objectives for FEATUR.UX. Given access to multitrack music and audio feature extractors, we hypothesize that the prototype affords users greater control of audio data and new methods to present visualisations. The technical load, reliance on amassed libraries of video assets, and pre- and post- processing requirements of traditional live visual practice and performance are eliminated by the use of computer generated methods of drawing. A modular GUI provides a flexible, adaptable workspace.

As shown in Fig. 5(A), users draw directed paths between nodes in a grid to create visualisations. This mapping model is inspired by the design and interactive interfaces of visual synths and offers a space for improvisation and creative spontaneity. The layered screen buffers to which the visuals are exported allow greater preferences for creating composites (synchronous and layered graphic visualisations). The limited palette of windows and menu options inspire the user to create with less.

3.2.1 UI Design of FEATUR.UX

The modular user interface, shown in Fig. 6, is composed of separate panels from which the user can, (A) start and stop audio playback, (B) view selected paths, (C) create directed paths within the workspace grid, (D) manually control playback of individual or groups of stems, (E) monitor the live waveform and spectrum, (F) view visualizations of the audio feature response to audio, (G) control which stem or group of stems is visualized in (F), (H) control which feature is visualized in (F), (I) manipulate parameters of graphic attributes, (J) monitor a log of which audio features are mapped to which visual attributes, and (K) view the composite visualizations created in the application.

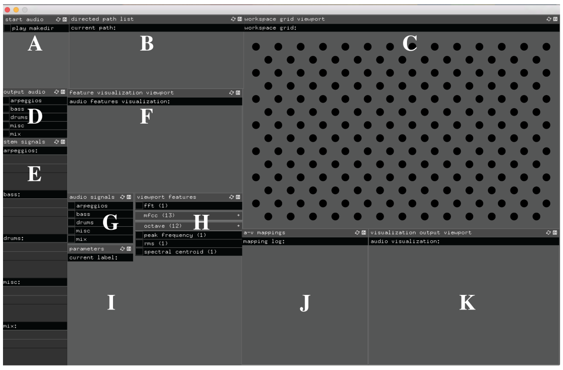

Figure 6. The modular GUI in FEATUR.UX.

3.2.2 Workspace Grid

The main workspace, shown in Fig. 6(C), is a grid within which the user can draw paths between pins. As shown in Fig. 7, a completed path starts with an input audio stream pin and ends with an output buffer window pin. The paths in between the input and output pins can include different combinations of pins that control color, thresholds and parameters for graphic methods of drawing.

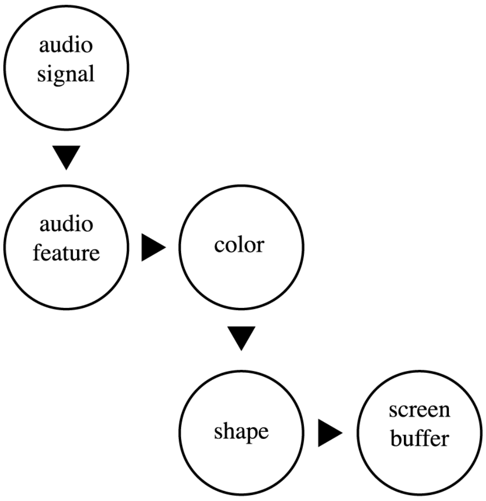

Figure 7. Data flow between pins.

One instance of an audio feature can be used to control the behavior of several visual parameters as shown in Fig. 8. One stem can be connected to multiple audio features, whose visualizations can be combined into a single layer in a shared screen buffer as shown in Fig. 9. And, many instances of any type can be used within a directed path as seen in Fig. 10. Also, if more than one audio stream is connected at the input pin of a path, the audio data used to generate the visualization is the mixed audio signal.

[12] http://openframeworks.cc/

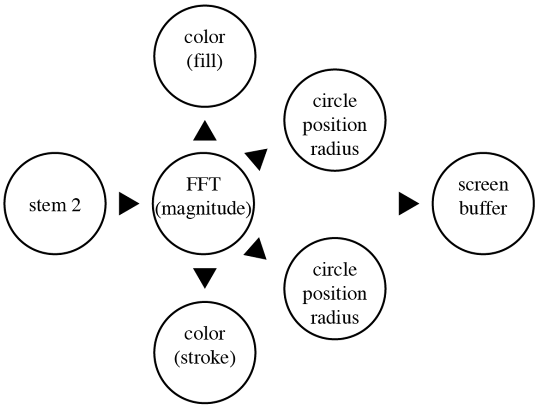

Figure 8. One stem and one audio feature is used in this path to control separate visual parameters as shown. A parameter of the FFT is used to manipulate the color, fill, stroke and position of circles as shown in the resulting layer composition in Fig. 11(A) and (B).

3.2.3 *Audio Input Panel*

There are three proposed cases in which multitrack audio can be imported into FEATUR.UX as seen in Fig. 4. In the case that the mixed audio track is not a sum of individual stems, the audio for the mixed track can be routed to the output channels while the user uses the data from the constituent stems to create visualisations. If the imported audio file is a Stem format file or the final mix is the sum of the separate stems (and the imported file is not a Stem format file), the user can control which audio streams are routed to the output channels while creating visualisations with the respective streams of data. Lastly, for live audio inputs, the audio for each stem is routed to the output channels while the user visualizes the live data.

3.2.4 *Feature Visualisation Panel*

FEATUR.UX lets the user monitor the response of audio features. The user selects a feature to visualize, as shown in Fig. 6(H), by choosing one or more audio stems as shown in Fig. 6(G). If more than one audio input is selected, the mixed audio is used to create the feature visualisation for the chosen extractor.

3.2.5 *Dynamic Parameter Panel*

Sections of this panel as shown in Fig. 6(I) appear only after a user selects a visual parameter pin along a closed path in the workspace grid. The FEATUR.UX interface is designed to offer access only to UI elements that are required to complete the task being considered.

3.2.6 *Output from Screen Buffer Window*

Each directed path in the workspace grid ends with a screen buffer as shown in Fig. 7. For every completed path, there exists a separate, layered visualisation in order of user preference, as shown in Fig. 11(A) through (F). Access to parameters to manipulate the appearance of the screen buffer are dynamically accessible as mentioned above. A composite image of a layered visualization is shown in Fig. 11(G).

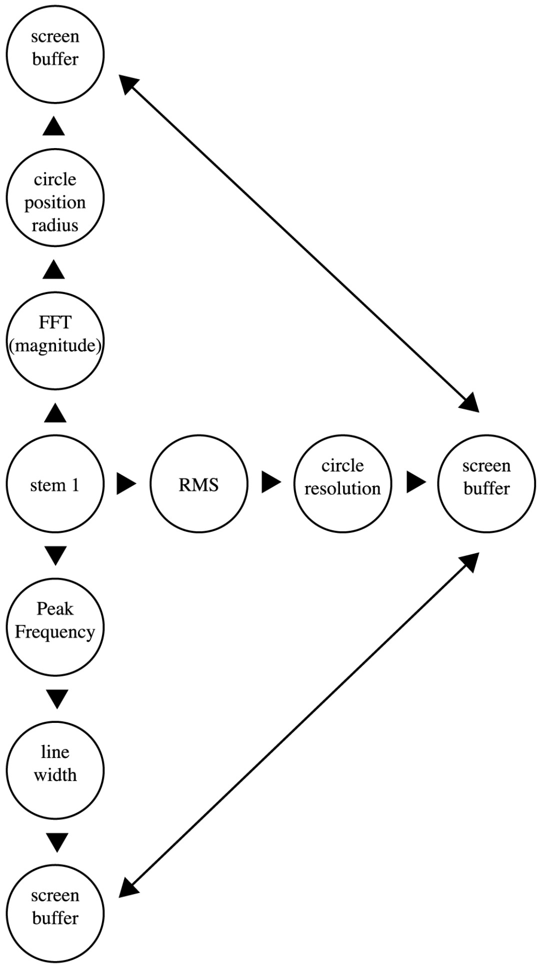

Figure 9. One stem and three audio features are used in this path to control separate visual parameters as shown in the resulting layer composition in Fig. 11(C) and (D).

3.2.7 *Feature Extraction in FEATUR.UX*

The following are the current audio features available in the FEATUR.UX prototype:

1. The Fast Fourier Transform extracts spectral information from an audio signal. The resulting complex signal is composed of a real and an imaginary part, which are used to calculate the magnitude and phase of the signal [30]. The FFT calculation performs as an auditory filter that mirrors, to some extent, the physiology within the human ear [31]. This perceptually-salient audio feature is a standard extractor used for sound visualisation.
2. MFCCs are a perceptual feature used to represent the timbral characteristics of an audio signal [32]. The representation of the short-term power spectrum is usually depicted using 8 - 20 of the first coefficients. The number of coefficients used can be adjusted based on the complexity of the signal [12]. Each of the coefficients can be isolated separately to monitor its behaviour and visualise.
3. The Chromagram, referred to as the Octave Analyzer in ofxMaxim, reveals the distribution of energy in an audio

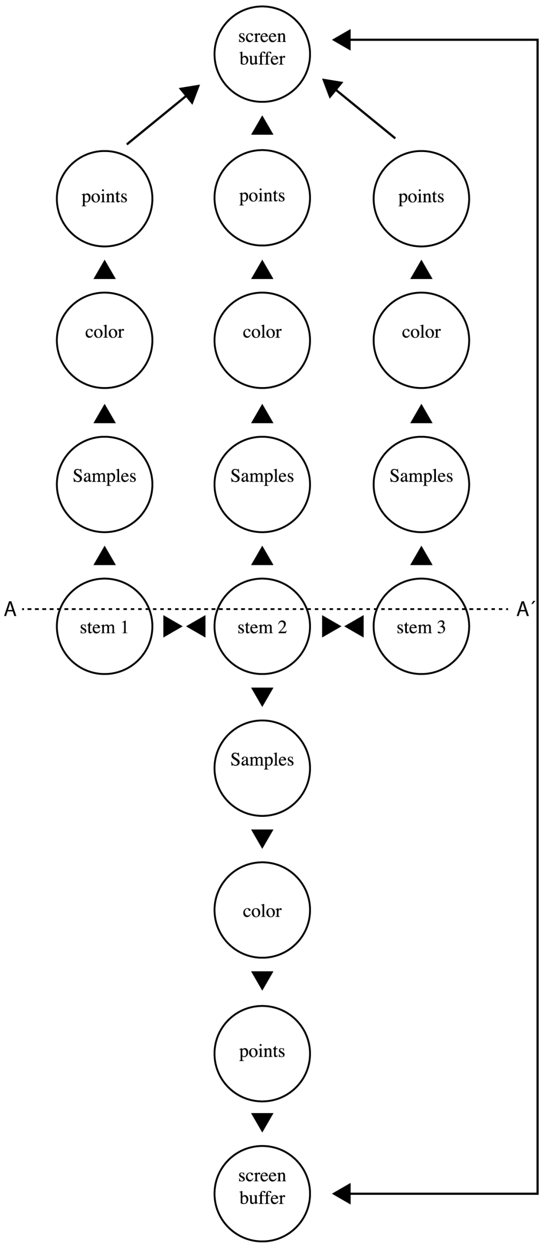

Figure 10. From the cross section A-A, downwards, the sample data is created with a mixed audio stream. Three stems are combined to influence the color and position of points. Upwards from the cross section A-A, shows how sample data from each stem can also be used to control the individual behavior of visual attributes. The visualisations that result from the featured path are shown in Fig. 11(E) and (F).

signal along a range of pitches. The dimension of tone height, where the range is segmented into octaves rather than pitch classes from traditional music scales [12] [21]. Each of the 12 pitch classes can be isolated separately to monitor its behaviour and visualise.

4. The waveform is an aggregate of compound sinusoidal waves and makes up the raw audio signal [32]. The waveform itself is not considered an audio feature, but is a ubiquitous method used to create sound visualisations.

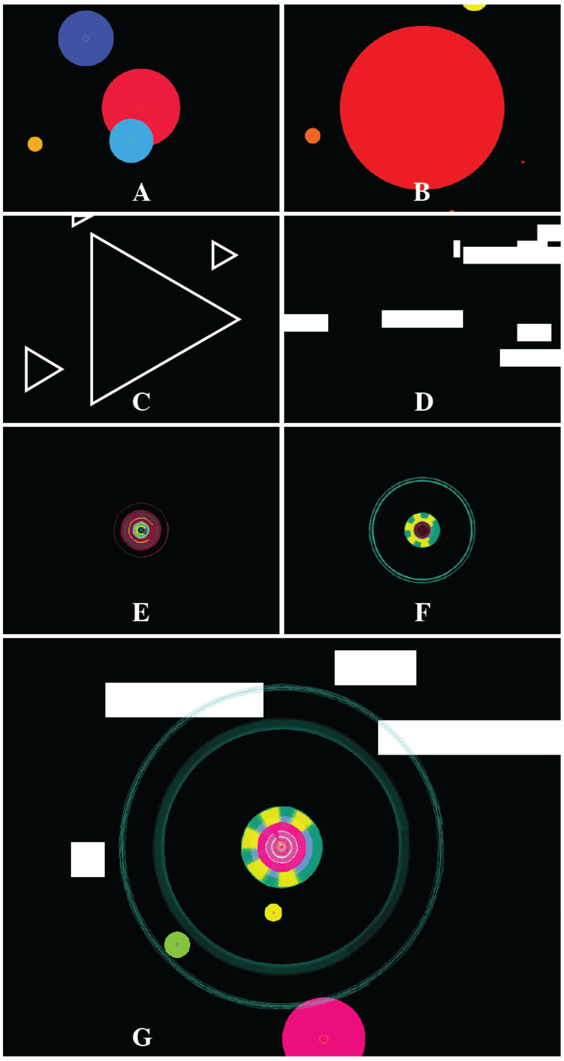

Figure 11. Images (A) and (B) display visualizations generated from the directed path shown in Fig. 8. The directed path shown in Fig. 9 generated the forms seen in images (C) and (D). The circle resolution is controlled by the RMS value, an indication of loudness. In image (C) where it is valued between 3.0 and 4.0 (The upper bound is exclusive), triangles are generated. In image (D), the RMS is at least 2.0 and at most 2.99, therefore a line is drawn. Larger RMS readings generate shapes that closer represent the circular form. Peak Frequency readings are expressed in the resulting line widths. The compound path shown in Fig. 10 is used to generate the meshes drawn in images (E) and (F) in which the graphic depicted in yellow represents the mixed audio stream. The meshes that represent the behaviour of the individual stems are rendered separately. A frame of a multi-layered visualization composed with the paths described in Fig. 8, Fig. 9 and Fig. 10 is shown in image (G). Wombatman6581 was used to generate this visualization. Musician Goto80 produced the song using a Commodore 64 with a 6581 SID-chip.

5. The peak frequency is the measure of the frequency bin with the highest magnitude within the spectrum of an audio signal. In some cases, it is an indicator of pitch, which may reveal the position of the fundamental frequency [33].

6. The spectral centroid is the frequency corresponding to the center of gravity of the energy spectrum. It is the threshold at which half of the energy is above or below that frequency. The measure of the spectral centroid relates to the perception of brightness or sharpness and quality of timbre that increases with the presence of high frequencies [12] [32].

7. The Root Mean Square (RMS) is a measure of signal intensity that evaluates the envelope of an audio signal and can be seen as a basic model of loudness of an audio signal [12] [32].

4. CONCLUSIONS AND FUTURE WORK

The adoption of multitrack audio for creative applications is still new and the use of stem technology is growing. [13] FEATURE.UX proposes to introduce a platform to apply multitrack audio towards live audio visual performance.

The quantitative results from the preliminary survey of a limited sample of participants reveal that introducing additional audio features and multitrack audio into the pipeline for developing live visuals is in its infancy. It is our assumption that the responses to the survey lacked clear motivations to use multitrack audio because there exists a lack of tools and opportunity to do so. With FEATURE.UX we aim to provide a framework to be able to test this hypothesis. The lack of awareness about audio features beyond the commonly exploited extractors and of multitrack stems is significant. Furthermore, the evaluation of the qualitative results thus far supports the earlier findings of Carvalho and Lund [24]. The qualitative results from the survey reveal that few participants use a common language to discuss topics related to live audiovisual practice and performance and, the departmentalization of the various disciplines within the audiovisual space creates a barrier that inhibits communication. Although at least 2 participants noted that they currently use stems in their audiovisual practice, the utility of multitrack audio visualization will remain unknown until it is experienced by more users. Additional studies will be conducted to learn how the community considers and implements mapping between sound and image and to further explore the use of audio features and stems to control parameters for generative computer visuals.

5. REFERENCES

[1] N. Correia Da Silva Diniz, M. Demey, and M. Leman, “An interactive framework for multilevel sonification,” 2010.

[2] H. J. Song and K. Beilharz, “Aesthetic and auditory enhancements for multi-stream information sonification,” in *Proceedings of the 3rd international conference on Digital Interactive Media in Entertainment and Arts*. ACM, 2008, pp. 224–231.

[3] S. Jordà, “The reactable: tangible and tabletop music performance,” in *CHI’10 Extended Abstracts on Human Factors in Computing Systems*. ACM, 2010, pp. 2989–2994.

[4] N. Corriea, *Interactive Audiovisual Objects*. Aalto ARTS Books, 2013.

[5] R. Hiraga, R. Mizaki, and I. Fujishiro, “Performance visualization: a new challenge to music through visualization,” in *Proceedings of the tenth ACM international conference on Multimedia*. ACM, 2002, pp. 239–242.

[6] R. Dahyot, G. Kearney, and C. Kelly, “Visual enhancement using multiple audio streams in live music performance,” in *Audio Engineering Society Conference: 31st International Conference: New Directions in High Resolution Audio*. Audio Engineering Society, 2007.

[7] E. H. Margulis, “A model of melodic expectation,” *Music Perception: An Interdisciplinary Journal*, vol. 22, no. 4, pp. 663–714, 2005.

[8] K. Brougher and O. Mattis, *Visual Music: Synaesthesia in Art and Music Since 1900*. Thames & Hudson, 2005.

[9] I. Bergstrom and R. B. Lotto, “Harnessing the enactive knowledge of musicians to allow the real-time performance of correlated music and computer graphics,” *Leonardo*, vol. 42, no. 1, pp. 92–93, 2009.

[10] G. Fazekas and M. Sandler, “Structural decomposition of recorded vocal performances and it’s application to intelligent audio editing,” in *Audio Engineering Society Convention 123*. Audio Engineering Society, 2007.

[11] S. Hargreaves, A. Klapuri, and M. Sandler, “Structural segmentation of multitrack audio,” *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 20, no. 10, pp. 2637–2647, 2012.

[12] D. Mitrović, M. Zeppelzauer, and C. Breiteneder, “Features for content-based audio retrieval,” *Advances in computers*, vol. 78, pp. 71–150, 2010.

[13] R. Stables, S. Enderby, B. De Man, G. Fazekas, and J. Reiss, “Safe: A system for the extraction and retrieval of semantic audio descriptors,” in *15th International Society for Music Information Retrieval Conference (ISMIR 2014)*, 2014.

[14] D. Smalley, “Spectromorphology: explaining soundshapes,” *Organised sound*, vol. 2, no. 02, pp. 107–126, 1997.

[15] T. Wishart, *On Sound Art*. Harwood Academic Publishers, 1996.

[16] N. Adams and M. Simoni, “Time-frequency visualization of electro-acoustic music,” in *Enriching Scholarship Workshop, University of Michigan*, 2005.

[17] J. Humiecka-Jakubowska, “Electronic music in the perspective of semiotics,” *Interdisciplinary Studies in Musicology*, no. 14, p. 258 273, 2014.

[13] http://www.stems-music.com/stems-partners/

[18] N. Valsamakis, “Aesthetics and techniques in the electroacoustic music of iannis xenakis,” 2000.

[19] A. Cameron, *Instrumental Visions: Electronica, Music Video, and the Environmental Interface*. Oxford University Press, 2013.

[20] C. Roads, *Microsound*. MIT Press, 2001.

[21] M. A. Casey, R. Veltkamp, M. Goto, M. Leman, C. Rhodes, and M. Slaney, “Content-based music information retrieval: Current directions and future challenges,” *Proceedings of the IEEE*, vol. 96, no. 4, pp. 668–696, 2008.

[22] K. Dahan, *Electroacoustic Music: Overcoming Analysis Paralysis*. Ann Arbor, MI: Michigan Publishing, University of Michigan Library, 2011.

[23] C. Baker, *VJam Theory*. Realtime Books, 2008.

[24] A. Carvalho and C. Lund, *The Audiovisual Breakthrough*. Fluctuating Images, 2015.

[25] J. Hook, D. Green, J. McCarthy, S. Taylor, P. Wright, and P. Olivier, “A vj centered exploration of expressive interaction,” in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 2011, pp. 1265–1274.

[26] N. N. Correia, A. Tanaka *et al.*, “User-centered design of a tool for interactive computer-generated audiovisuals,” in *Proc. of 2nd Int. Conf. on Live Interfaces*, 2014.

[27] M. Faulkner, *VJ: Audio-Visual Art and VJ Culture*. Laurence King Publishing, Ltd., 2006.

[28] S. Callear, *Audiovisual Particles: Parameter Mapping as a Framework for Audiovisual Composition*. Bath Spa University, 2012. [Online]. Available: https://books.google.co.uk/books?id=TTvYoAEACAAJ

[29] M. Grierson, “Maximilian: A cross platform c++ audio synthesis library for artists learning to program,” in *Proceedings of the International Computer Music Conference, New York*, 2010.

[30] S. W. Smith, *The Scientist and Engineers Guide to Digital Signal Processing*. California Technical Publishing, Ltd., 1997.

[31] S. Ravindran, K. Schlemmer, and D. V. Anderson, “A physiologically inspired method for audio classification,” *EURASIP Journal on Advances in Signal Processing*, vol. 2005, no. 9, pp. 1–8, 2005.

[32] A. Lerch, *An Introduction to Audio Content Analysis*. John Wiley and Sons, Inc., 2012.

[33] O. Tchernichovski, F. Nottebohm, C. E. Ho, B. Pesaran, and P. P. Mitra, “A procedure for an automated measurement of song similarity,” *Animal Behaviour*, vol. 59, no. 6, pp. 1167–1176, 2000.